Improvement in USB
Range Rate Accuracy
by
Rejection of Contaminated Data

1 December 1966

Contract NAS 5-3743

Prepared for
GODDARD SPACE FLIGHT CENTER
Greenbelt, Maryland

by

ELECTRO-MECHANICAL RESEARCH, INC.
AEROSPACE SCIENCES DIVISION
College Park, Maryland

INTRODUCTION

This paper contains a further statistical analysis of simulated USB static doppler tracking data. The use of the GSFC Apollo Centralized Computer System, the cooperation of station personnel at Guam, and the initial establishment of test parameters were all arranged by George Q. Clark of the Manned Flight Support Office. The data under consideration were received from Guam on July 7, 1966. The transmitter output was translated in frequency and attenuated so that the resulting signal effectively simulated zero doppler frequency. The doppler measurements were made in the destruct mode at a sampling rate of 10 per second and recorded in units of cm. per second. The receiver bandwidth was 50 cps throughout this test.

The data were taken in a number of "runs". Each run consisted of several samples of observations: the first taken at -132dbm (approximately 16 db above threshold), the second at -134dbm, etc. in steps of 2 db until threshold was reached. Then observations were taken in the same steps of 2 db back up to -132 dbm. The number of observations in each step varied from 30 to 1000.

The large amount of data provided an opportunity to investigate empirically the underlying probability distribution of the doppler data. Probability plots were used to determine how well the data followed a Gaussian distribution. As expected from previous work [1] for high S/N the distribution is very nearly Gaussian, but for lower S/N it departs significantly from it. In the light of the findings a model based on a contaminated Gaussian distribution is suggested to describe the distribution at the lower S/N, and the problem of estimating the mean (bias) and dispersion or precision of the observations is discussed for this model.

A GRAPHICAL TEST FOR NORMALITY

PROBABILITY PLOTS

A simple but effective graphical method to determine whether a random sample of measurements, $x_1, x_2, \ldots, x_n$ departs from a Gaussian distribution,

$$f(x{:}\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], -\infty < x < \infty,$$

is to plot the sample on normal probability paper. The sample is first ordered from smallest to largest value, letting $x_{(1)}$ be the smallest value, $x_{(2)}$ the next smallest, etc. The ordered sample values,

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

are called the order statistics of the sample.

The i^{th} order statistic, $x_{(i)}$, is plotted on the vertical axis of normal probability paper vs $\frac{i}{n+1}$ on the horizontal axis for $i = 1, \ldots, n$. The horizontal axis of the graph paper is scaled from 0 to 1 by stretching out the tails near 0 and 1 so that a reasonably "good" sample from a Gaussian distribution plots more or less along a straight line. Because of the randomness of real measurements, the points from a sample will deviate from the theoretical straight line, but for large samples, the linear tendency is unmistakable. For example, the readings at -132 dbm (Step 0) for Run 2 of the July 7th data as plotted in Figure 3 clearly seem to come from a distribution that is very close to Gaussian.

Departures from normality show up in probability plots as non-linearities. For example, if the sample is from a rectangular distribution (dashed line, Figure 1) instead of a Gaussian (solid line), the plot will resemble an "S" curve (dashed line, Figure 2) instead of the straight line for the Gaussian since the extreme

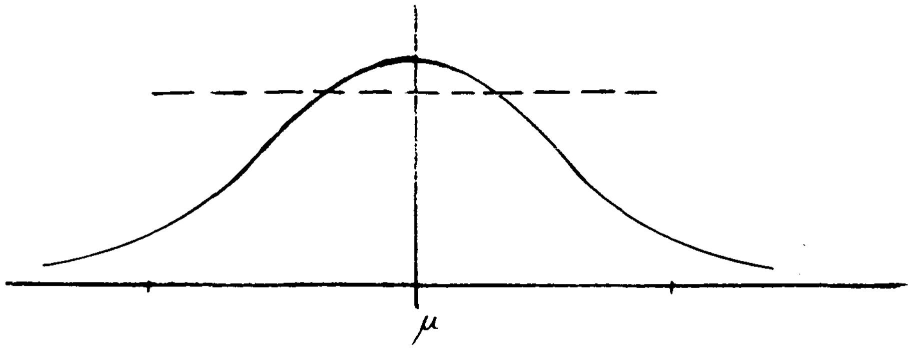


Figure 1

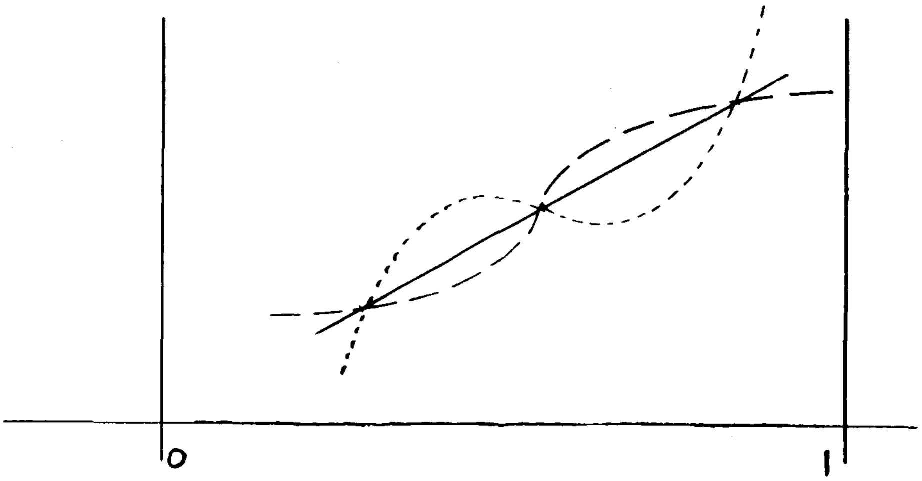


Figure 2

measurements are not as far out as they would be for the Gaussian. On the other hand, if the sample comes from a distribution with more probability on the tails than a normal, for example, a Cauchy distribution,

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \mu)^2} , \quad -\infty < x < \infty$$

The plot will have a reversed "S" shape (dotted line, Figure 2). Although the Cauchy density function looks very much like the normal curve when graphed, there is considerably more probability way out on the tails so that the chance of extremely large or small readings is much greater.

Hence, the reversed "S" shape of the probability plots.

CONTAMINATION MODELS

Quite often sample data is basically Gaussian except that it is contaminated by a few (or many) measurements from some other distribution. That is, some of the measurements are influenced by unusual errors caused by anything from bursts of noise in a circuit to blunders in reading a dial. These measurements are called contaminants and carry misleading information about the value being estimated. Sometimes they show up as outliers, i.e. outrageously large or small readings, which can simply be discarded before using the data. More often they are mixed in with the good data and are difficult to identify.

One simple contamination model is given by the density function

$$f_c(x) = (1-\gamma)\frac{1}{\sqrt{2\pi}\,\sigma_o}\exp\left[\frac{-(x-u)^2}{2\sigma_o^2}\right] + (\gamma)\frac{1}{\sqrt{2\pi}\,\sigma_1}\exp\left[\frac{-(x\mu)^2}{2\sigma_1^2}\right],$$

$$-\infty < x < \infty,$$

where $0 < \gamma < 1$, $\sigma_1 >> \sigma_o$. In this model it is assumed that 100 $(1-\gamma)$ per cent of the measurements are "good", i.e. come from the primary Gaussian distribution with standard deviation σ_o, but 100 γ per cent of them come from a contaminating distribution, here also Gaussian, but with a very much larger standard deviation σ_1.

For example, if $\gamma = .01$ and $\sigma_1 = 10\sigma_o$, a sample of size 100 from this distribution typically will appear to be quite normal except perhaps 1 or 2 extremely large or small measurements which obviously should be discarded before computing estimates of μ and σ_o. However, if $\gamma = .10$ and $\sigma_1 = 3\sigma_o$, the problem is more difficult since the contaminants are frequent and well mixed in with the "good" data.

The model above is an example of scale contamination because the contaminating distribution has the same mean as the primary distribution, but it has a much larger standard deviation. Other models of scale contamination with non-normal or even unsymmetric contaminations are possible.

PROBABILITY PLOTS OF THE JULY 7 DOPPLER DATA

Probability plots of some of the July 7 data are given in Figures 3 through 8. Samples of about 500 readings are plotted for S/N from -132 dbm (Step 0) to -146 dbm (Step 7). On each graph the point $(x_{(i)}, \frac{i}{n+1})$ is labeled " i " for identification. Not all of the points are plotted because of the lack of space on the graph, but, for example, in Figure 5 the points corresponding to $x_{(51)}, x_{(52)}, \ldots, x_{(99)}$ would fall about on a straight line between the points for $x_{(50)}$, $x_{(100)}$. Also, the graph must be monotonically increasing.

At Step 0 (Figure 4) the plot is still remarkably straight except for three outliers $x_{(1)}$, $x_{(498)}$, and $x_{(499)}$. By Step 4 (Figure 5) the points from $x_{(25)}$ to $x_{(498)}$ appear to be from a Gaussian distribution. At Step 6 (Figure 6) about the smallest 4 per cent and the largest 8 per cent of the readings are not from the Gaussian distribution which seems to characterize the rest of them. This tendency increases with decreasing signal level so that by Step 7 (Figure 8) only about the middle 50 per cent seem to be from the primary distribution. Even then, that middle 50 per cent plots very straight.

Thus, as the S/N decreases the linear shape gives way to the inverted "S" shape characteristic of a contaminated Gaussian discussed in the previous section. To illustrate, a theoretical curve for a scale contaminated Gaussian distribution with $\gamma = .08$, $\sigma_o = 14$, and $\sigma_1 = 84$ is drawn in Figure 9. Compare this curve with the empirical plot in Figure 6 and notice the similarity.

ESTIMATION OF UNKNOWN PARAMETERS FROM PROBABILITY PLOTS

As described above, probability plots of samples from a Gaussian distribution tend to follow a straight line. The intercept and the slope of this theoretical line are determined by the value of the mean μ and the standard deviation σ, respectively, of the distribution. The ordinate at which the line crosses the 50 per cent abscissa line will be the mean. Also, the larger the σ, the steeper the slope.

Accordingly, unknown values of μ and σ may be estimated from empirical probability plots of the sample data. A line of best fit is drawn to the data, either by least squares or some other method. The intercept of the 50 per cent abscissa line serves as an estimate of μ, and the slope of the line when converted to proper units is the estimate of σ. These graphical estimates based on least square curve fits have been thoroughly studied and are known to have useful optimality properties [2].

In Figure 3 a straight line is drawn in to fit the data from Step 0. This line was drawn simply by eye, but it is good enough to illustrate the ideas. The graphical estimate of σ from this line is $\tilde{\sigma} = 5.6$ whereas the root mean square estimate of σ computed from the same data is $\hat{\sigma} = 5.479$. Whenever the distribution is Gaussian the computed and the graphical estimates should agree fairly well.

For lower S/N the observations appear to follow the contaminated Gaussian model, and a line of best fit to all of the points is no longer meaningful. However, since the center portion of each sample does seem to follow a Gaussian law (the primary distribution referred to above), the tails can be ignored and a line drawn to fit that central portion of the plots in order to estimate the μ and σ_o of the primary distribution. These are not influenced by the tails of the sample and give estimates of the parameters with little influence by contaminants.

In Figure 10 are plotted the root mean square estimates of σ computed from the samples for the various steps in the second run of July 7. These are plotted with ". 's". Plotted by "x's" are the graphical estimates. As expected the graphical estimates are much lower than the computed root mean square estimates for the lower S/N.

At this juncture several conclusions should be noted.

1) For low S/N the usual estimates of μ and σ computed from the data, i.e. $\overline{x}$=sample mean and s=sample root mean square error, can be inefficient and misleading. $\overline{x}$ is strongly affected by a wild measurement, especially when the sample contains a small number of readings, and can easily give a bad estimated value. For example, if five readings are to be averaged, then one wild contaminant among the five can overshadow the other four.

Also, if s is used to estimate the scatter of the data (and hence the precision with which μ is known) wild readings will make it appear that less precision is possible than really is. From the probability plots it is clear that a certain portion of the readings are from the primary Gaussian distribution which has a much smaller standard deviation, σ_o, than the value s gives. If these readings can be weighted more heavily in estimating μ than the contaminants are, then the standard error of estimation would be closer to $\sigma_o/\sqrt{n}$ than to $s/\sqrt{n}$. Thus, it should be possible, by the use of carefully chosen estimates of μ, to know its value with better precision than the value of s indicates.

2) On the other hand, the precision possible in estimating μ will never be as good as if there were no contaminants. This is clear for two reasons. First, the contaminants are essentially wasted measurements, and second, even more information is lost because one does not know which of the readings are "good" and which are "bad".

To summarize these two points with an example consider Step 6. The r.m.s. estimate, s, is about 82. From this it follows the standard error of $\overline{x}$ as an estimate of μ will be about $82/\sqrt{n}$, where n is the number of readings to be averaged for $\overline{x}$. However, the graphical estimate of the standard deviation of the primary Gaussian distribution is about 20. Thus, if there were no contamination, $\overline{x}$ would have a standard error of about $20/\sqrt{n}$, about four times better than $82/\sqrt{n}$. The precision with which μ can actually be estimated will be less than $82/\sqrt{n}$ but greater than $20/\sqrt{n}$. How close it is to $2\sigma/\sqrt{n}$ depends on γ, the proportion of contaminants, and the ratio σ_1/σ_o.

3) Hence, the problem is to find an estimate of μ which makes reasonably efficient use of the information in the readings yet which is not seriously disturbed by contamination. A short discussion of the statistical principles involved is given in the next section.

ROBUST ESTIMATION

If a sample of size n, $x_1, \ldots, x_n$ from a Gaussian distribution with mean μ, variance σ^2 is to be used to estimate the unknown value of μ, then the sample mean, $\overline{x}$, has no serious competitor. Any other sample statistic computed from the data to estimate μ (e.g. sample median, sample midrange) is relatively inefficient, i.e., wasteful of information.

If, however, the sample is from another distribution, for instance a Cauchy (which looks very much like a normal curve when graphed), then $\overline{x}$ is not a good estimate of μ. In any sample of measurements from a Cauchy it is likely that there is a measurement so wild that it completely dominates the rest causing $\overline{x}$ to be far from the true value of μ. For the Cauchy distribution the sample median is a far better estimate of μ than $\overline{x}$ since it is not as influenced by extremely large or small readings. To a lesser extent the same holds true for the scale contaminated Gaussian model discussed above.

Thus, the worth of any sample statistic as an estimation of a parameter depends on what distribution is being sampled, and a given sample statistic may be the best possible in one case and almost worthless in another. If a statistic or estimate is useful over a range of distributions, it is said to be robust. For example, although the sample median is not the best for either the normal or the Cauchy, it is reasonably good for either and hence is robust over these two distributions. By using the median instead of $\overline{x}$ one trades some efficiency in sampling from Gaussian for safety in case the distribution is not really Gaussian but perhaps Cauchy-like instead.

With the doppler data under consideration the estimation problem seems to be that the bias, μ, must be estimated from data that is sometimes very nearly Gaussian (high S/N) but at other times may be like a scale contaminated Gaussian (lower S/N). To be useful an estimate of μ must have two properties:

a) If the data is Gaussian, the estimate must be reasonably efficient. That is, there must be no serious waste of information when compared with the best estimate $\overline{x}$.

b) If the data is contaminated Gaussian, the estimate must be relatively uninfluenced by wild measurements. In this case it must be sufficiently superior to $\overline{x}$ to justify the loss of efficiency in a).

Several types of estimates which possess these properties are the truncated means, Winsorized means, and the Hodges-Lehman statistics [3], [4], [5], [6].

Truncated means are computed by censoring extreme sample values before averaging. For example, suppose the orbit determination program requires an input every second and readings in the non-destruct mode can be taken as often as every 1/10th second. One way to specify the input to the program would be to use $\overline{x}$, i.e., take the ten 1/10th second readings and average them. (This is equivalent to taking just one 1 second reading.) Another method would be to take the ten 1/10th second readings, discard the smallest and the largest and average the middle eight. This is called a 10 per cent truncated mean, $\overline{x}_{(.10)}$, (i.e., 10 per cent of the observations on each end are censored). If the contamination is severe, then the average of the middle six observations, the 20 per cent truncated mean $\overline{x}_{(.20)}$, could be used.

It is well known that $\overline{x}$ is the most efficient estimate for the mean of a

Gaussian distribution for the case of uncontaminated data, but if there is some contamination, then $\overline{x}$ loses efficiency and the truncated means are better. Figure 11 is a graph of the asymptotic (large sample) efficiencies of means with various amounts of truncation. The graph, reproduced here from [3], is based on a scale contaminated Gaussian model with $\sigma_1 = 3\sigma_o$. Truncated means for finite sample sizes will exhibit similar characteristics.

Notice that even when the data is not contaminated, $\gamma = 0$, surprisingly little efficiency is lost by using a truncated mean instead of $\overline{x}$. From the graph it can be seen that the asymptotic efficiency of $\overline{x}_{(.06)}$ is about 97 per cent in the non-contaminated case. This efficiency can be interpreted to mean that as much information is obtained about the unknown parameter μ from 97 observations using $\overline{x}$ as from 100 observations using $\overline{x}_{(.06)}$.

Therefore, very little efficiency is lost by using these "safe" estimates when they are not really needed. However, when contamination is present, it can be seen from Figure 11 that the truncated means will perform much better than $\overline{x}$. With only 3 per cent contamination $\overline{x}_{(.06)}$ becomes as efficient as $\overline{x}$. With about 9 per cent contamination even the median is as good as $\overline{x}$.

The graph in Figure 11 is for $\sigma_1 = 3\sigma_o$. In the doppler data at Step 6 the amount of contamination is greater and σ_1 is approximately equal to $6\sigma_o$. For this case the efficiency of $\overline{x}$ will fall off still more rapidly than in Figure 11. Generally, the more severe the contamination, (i.e., either γ larger, σ_1/σ_o larger, or both) the greater the amount of censoring that should be used. This is done, of course, at the price of decreased efficiency in the noncontaminated case.

The Winsorized means, and more generally, the estimates based on weighted linear combinations of the order statistics, along with the Hodges-Lehman type statistic are a little more complicated to explain than the truncated means but exhibit similar useful properties. Just which one is the best depends on several factors, especially how much and what kind of contamination (or other non-Gaussian data) is expected, and what the cost or loss functions are for bad estimates under the various conditions. If, for example, the loss of efficiency in the high S/N case is of minor concern compared to the cost of a bad estimate at lower S/N, then the 20 per cent truncated mean will be preferred to the 10 per cent truncated mean.

SUMMARY AND CONCLUSION

In this report samples of simulated doppler tracking data with zero bias were analyzed to determine characteristics of the error distribution. It was found that, although for high S/N the error distributions are nearly Gaussian, for lower S/N they depart significantly from the Gaussian and follow more closely a scale contaminated Gaussian. The model could be refined further, but the use of the scale contaminated Gaussian suffices to determine the relative merit of various estimates of the bias.

The usual estimate, the sample mean, is not recommended because it has a severe loss of efficiency in the presence of scale contamination. Several other types of estimates which are known to be robust for contaminated distributions are suggested.

Specific recommendations as to the best estimates depend on the following items:

1) The data used in this report were obtained on one occasion from one tracking station. Before any application of the model suggested herein can be made, it is necessary to obtain data from other stations on other occasions to see if the characteristics of this data are typical.

2) The requirements of the orbit determination program will set restrictions on what kind of estimates can be used. The number of doppler data samples available to provide one input to the program will strongly influence the choice of estimate.

3) A knowledge of the loss or cost functions associated with the doppler estimation is needed so that an optimum estimator may be determined for all S/N ratios that are likely to be encountered during the actual mission.

Run 2, Step=0   n=504   -132 dbm

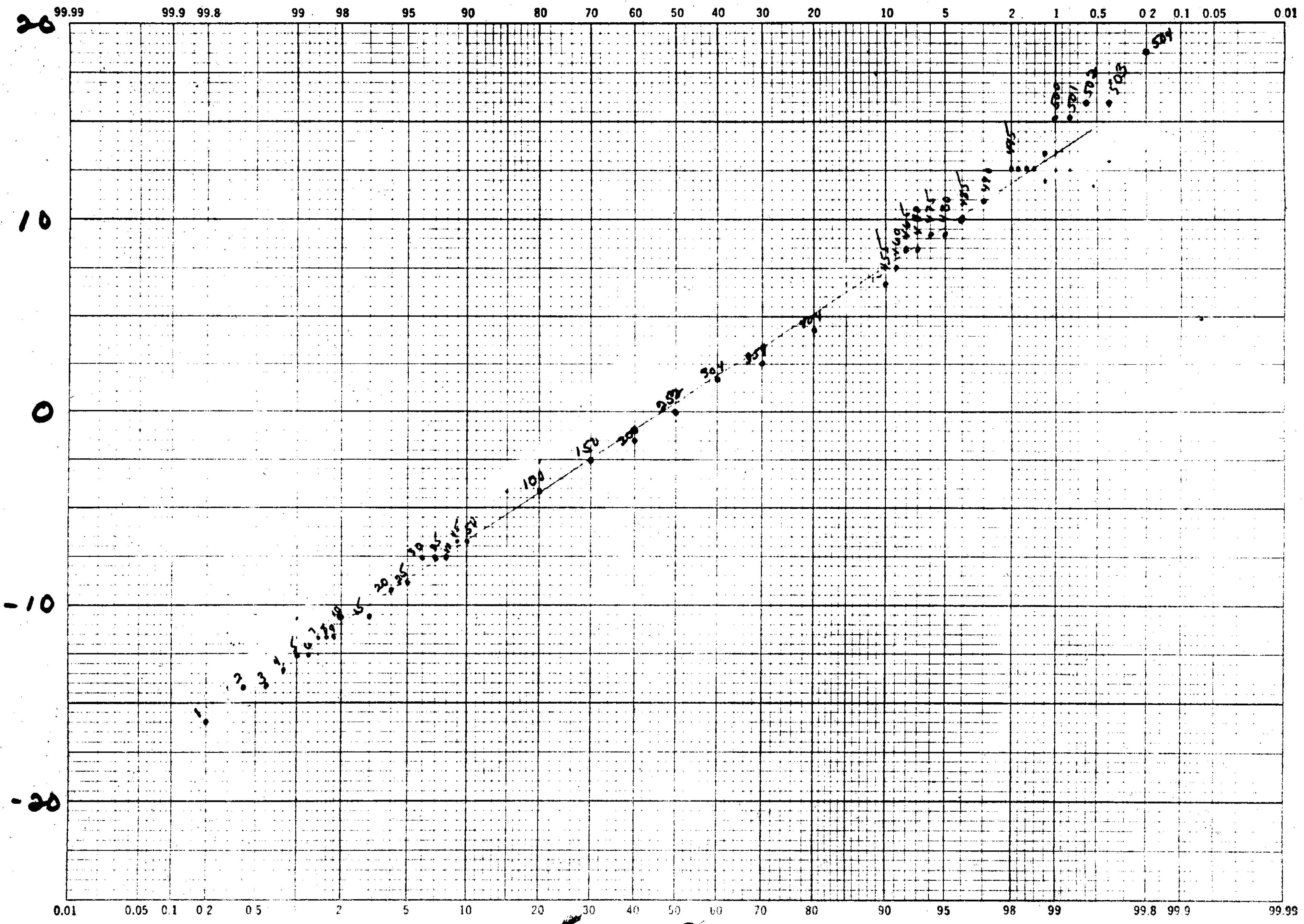


FIGURE 3

Run 2, Step = 2   n = 499   - 136 dbm

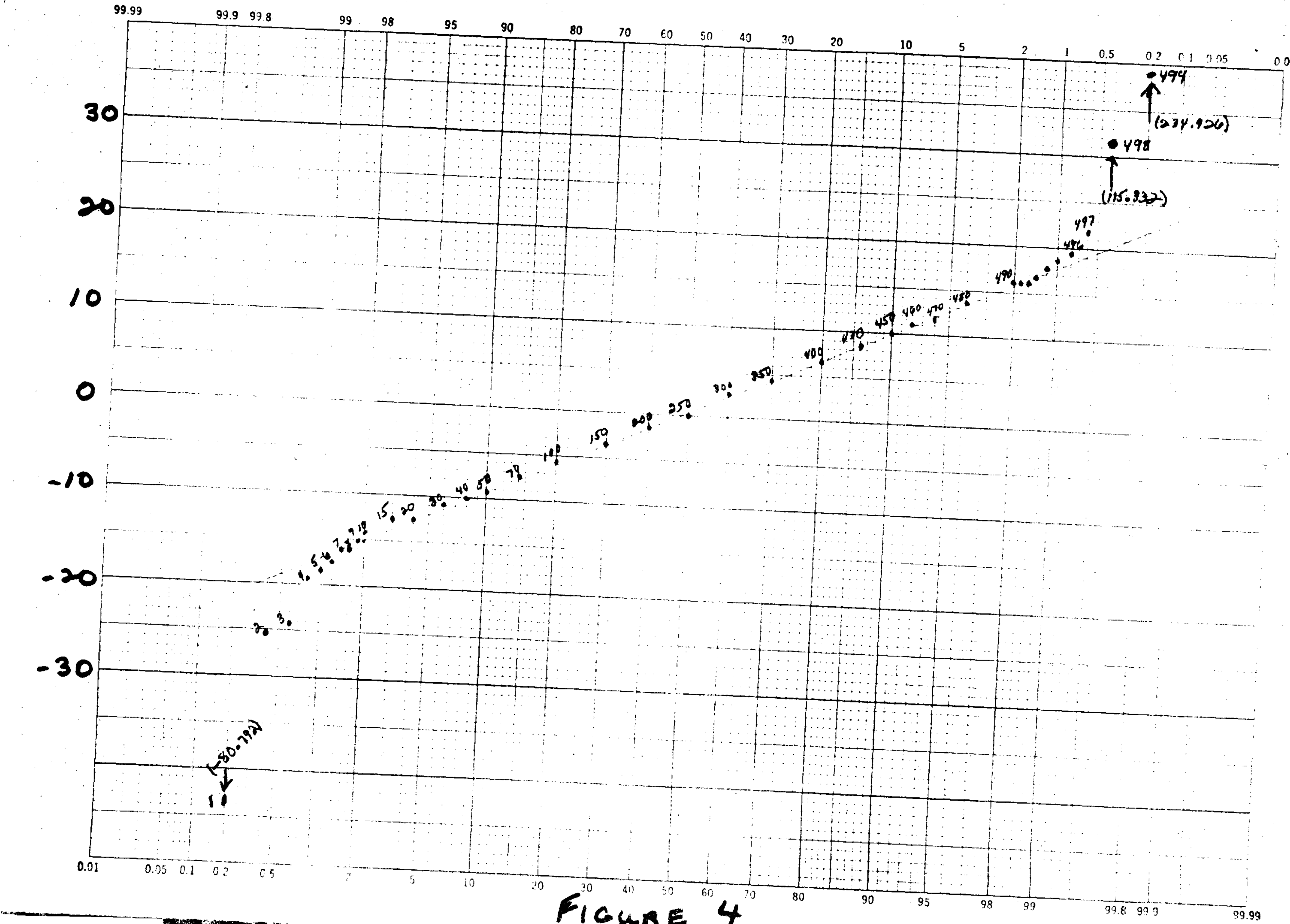


FIGURE 4

Run 2, Step = 4   n = 499   -140 dbm

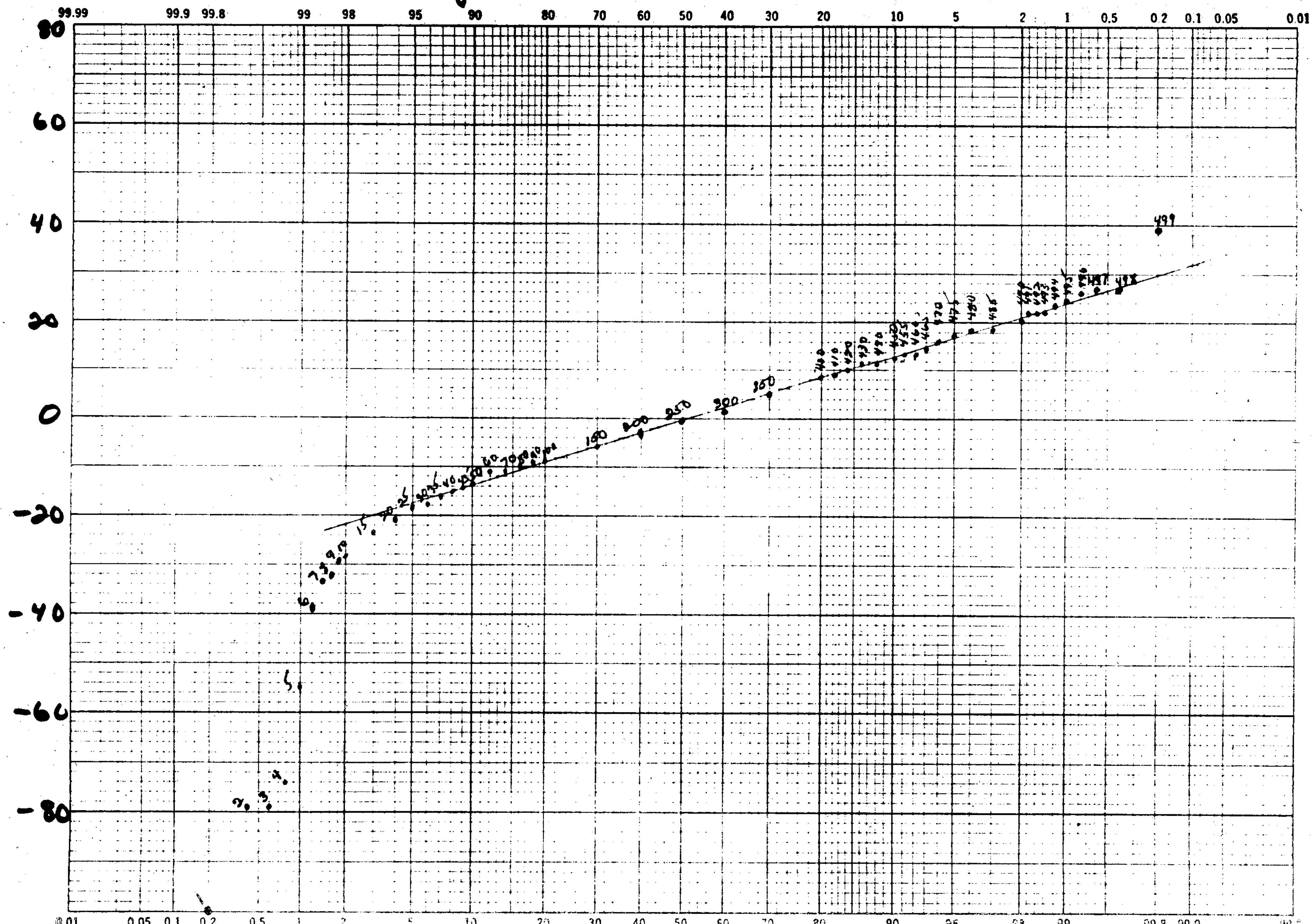
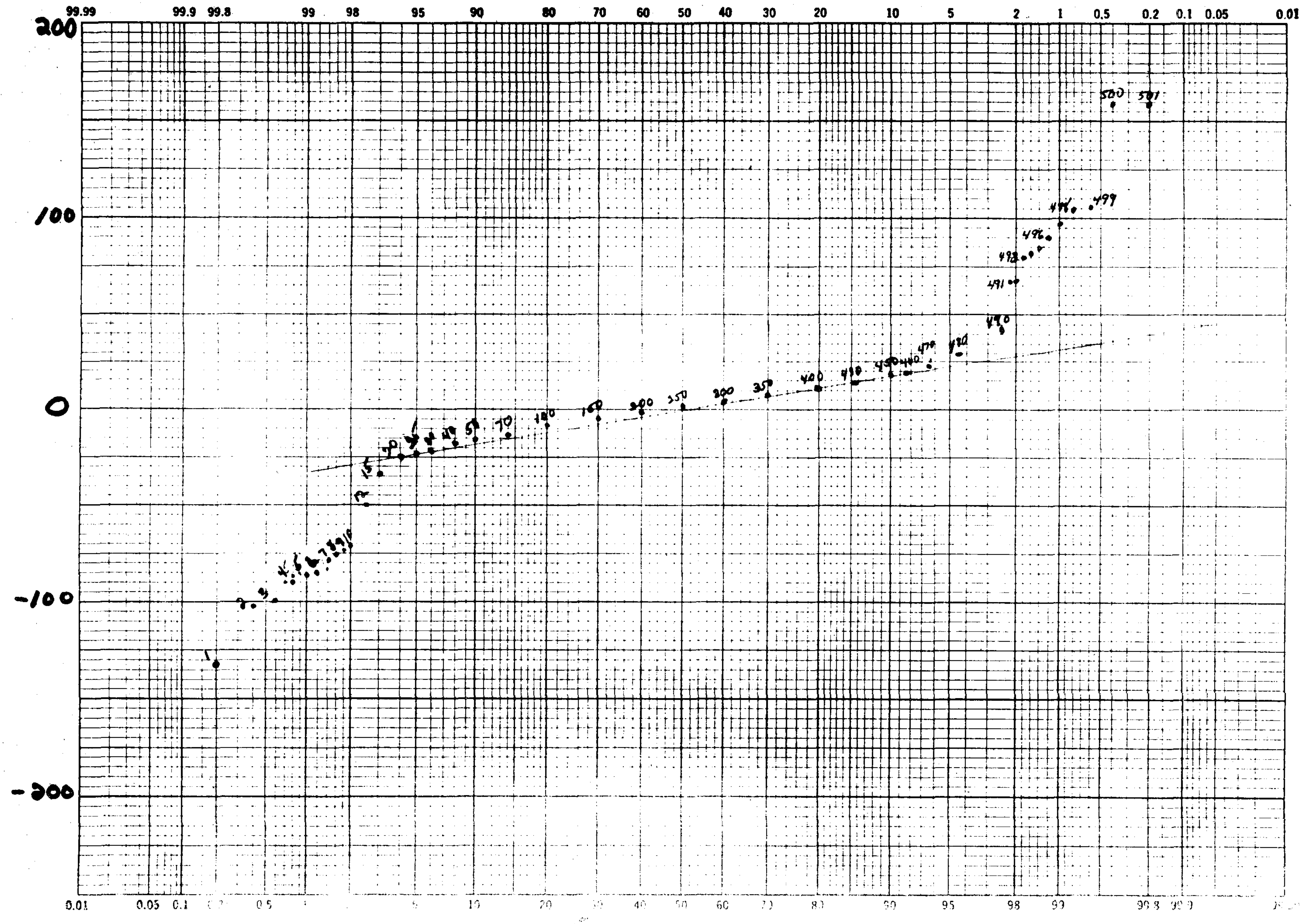
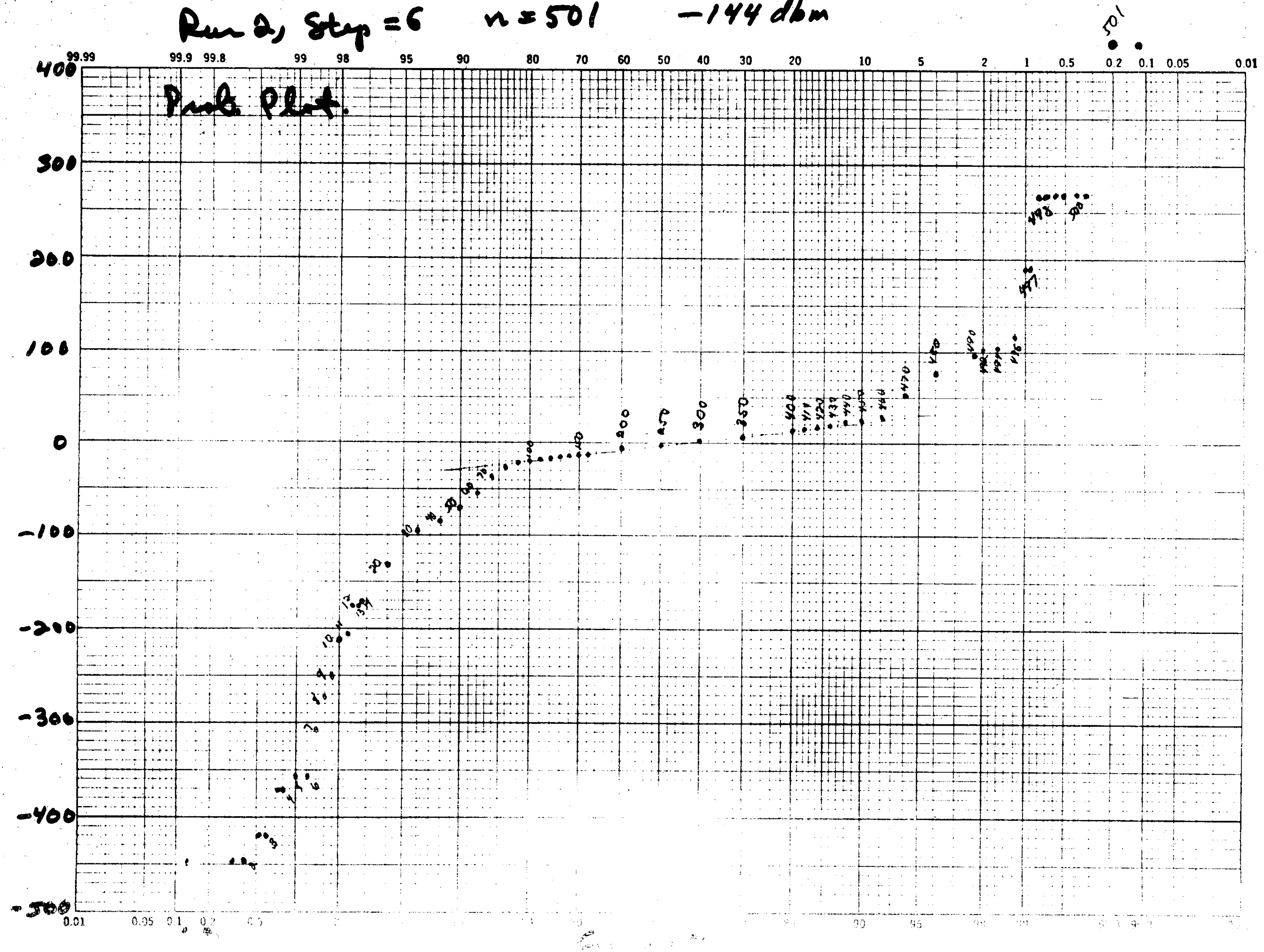


Figure 5

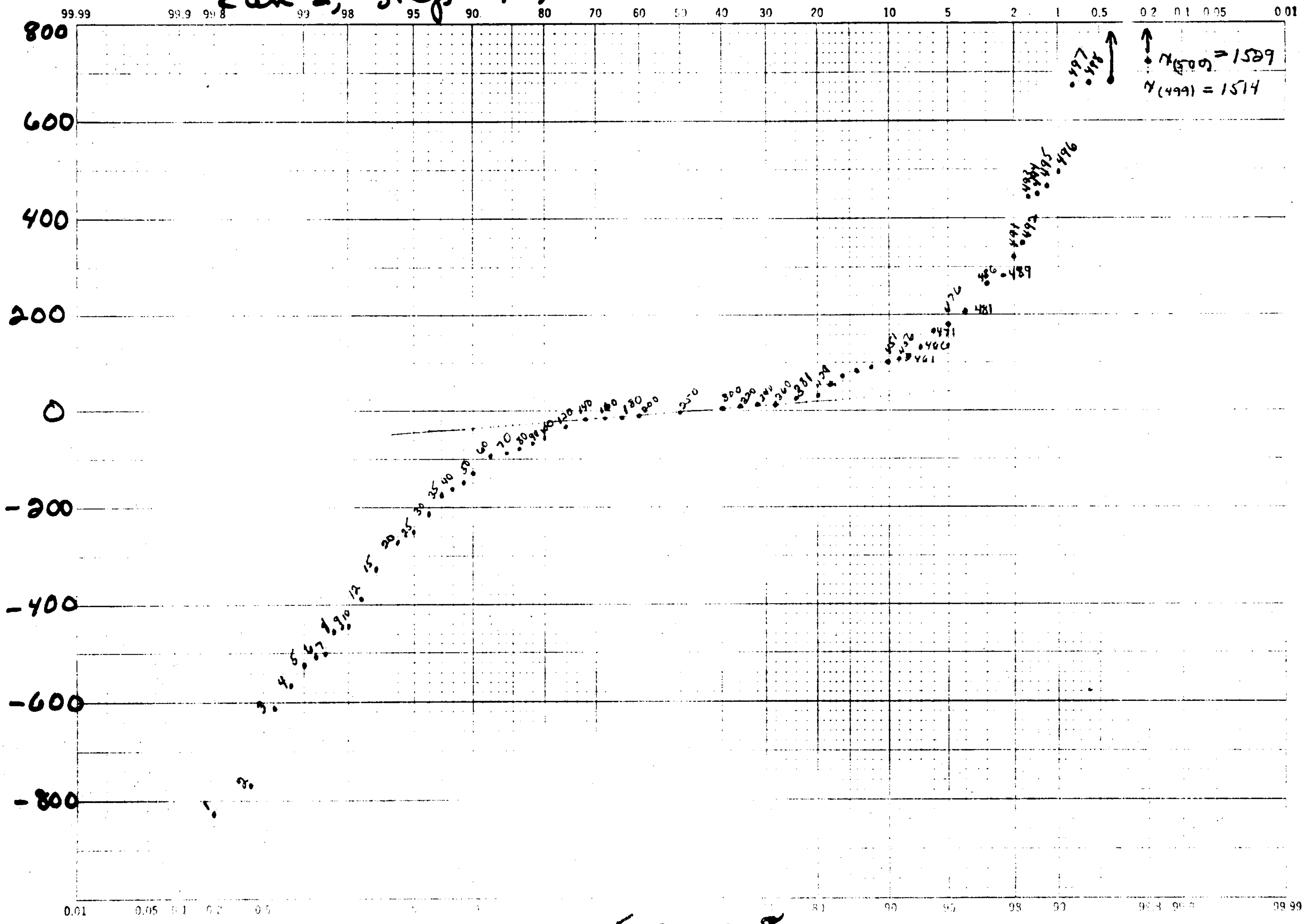
Run 2, Step=5 n=501 -142 dbm



Run 2, Step = 6   n = 501   -144 dbm



Run 2, Step = 7 , n = 500   - 146 dbm



FIGARE 8

THEORETICAL PROBABILITY PLOT: Contaminated Gaussian Distribution $\gamma = .08$, $\sigma_0 = 14$ $\sigma_1 = 84$

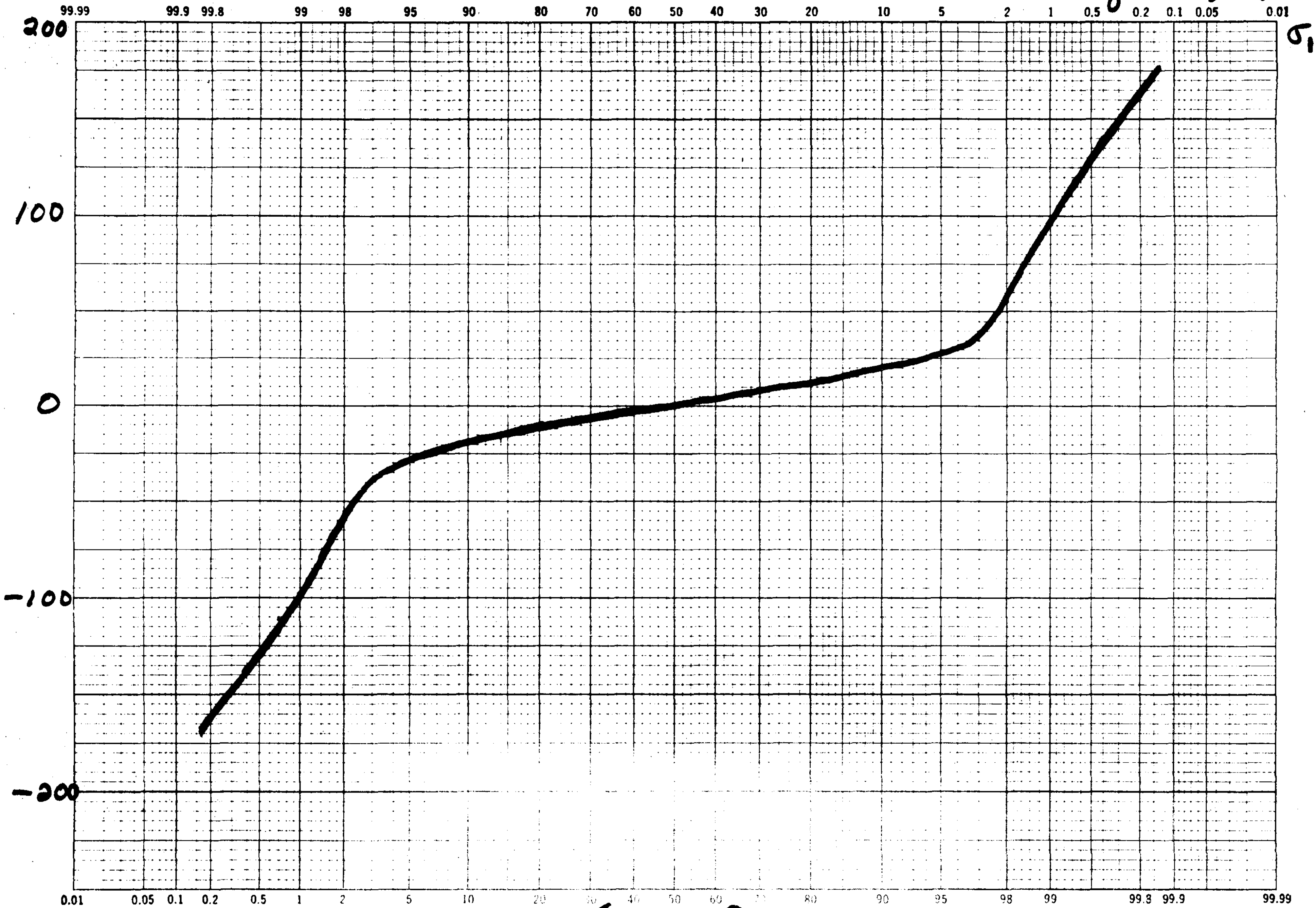


FIGURE 9

ESTIMATES OF σ, JULY 7 DATA

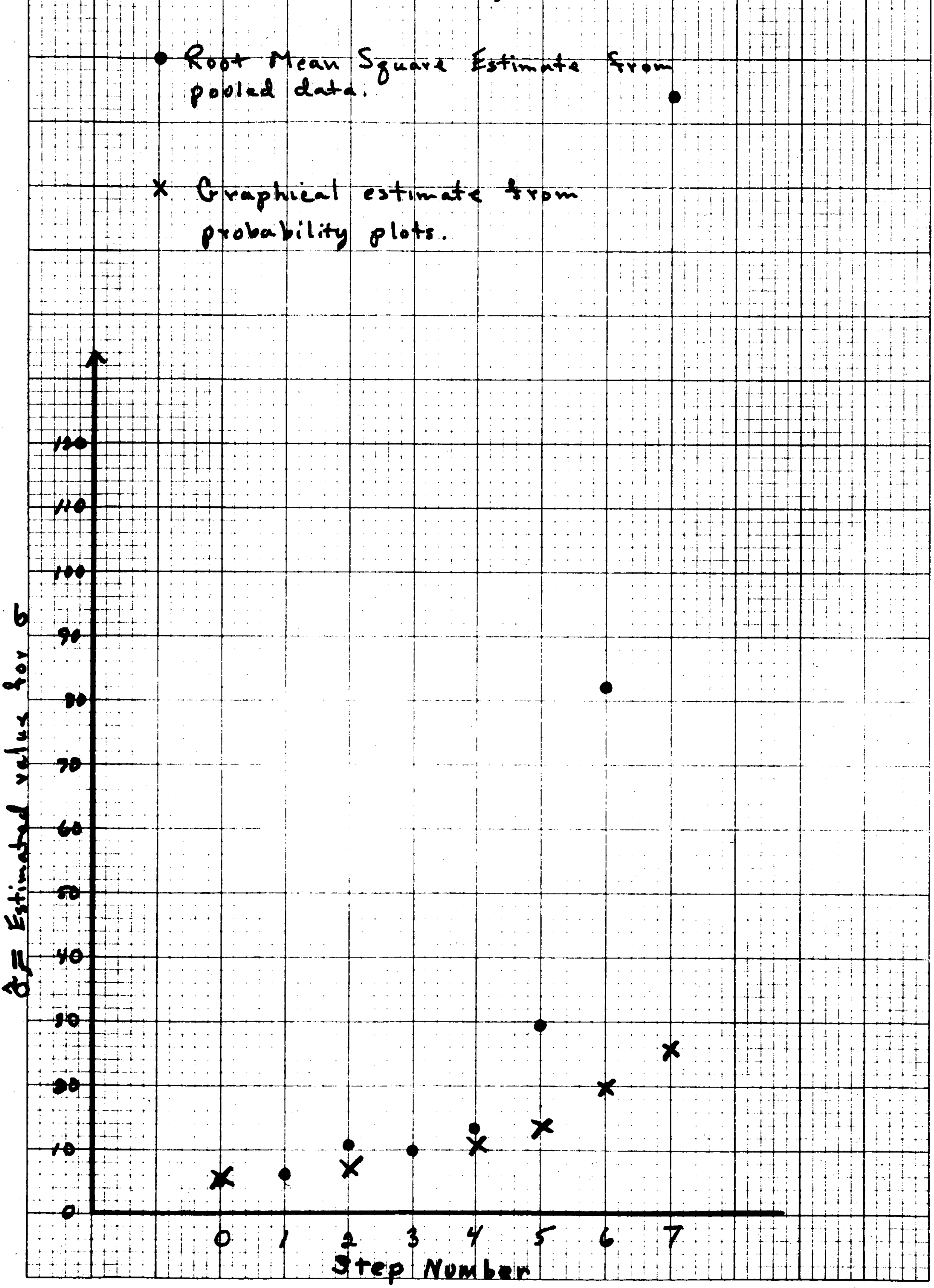


FIGURE 10

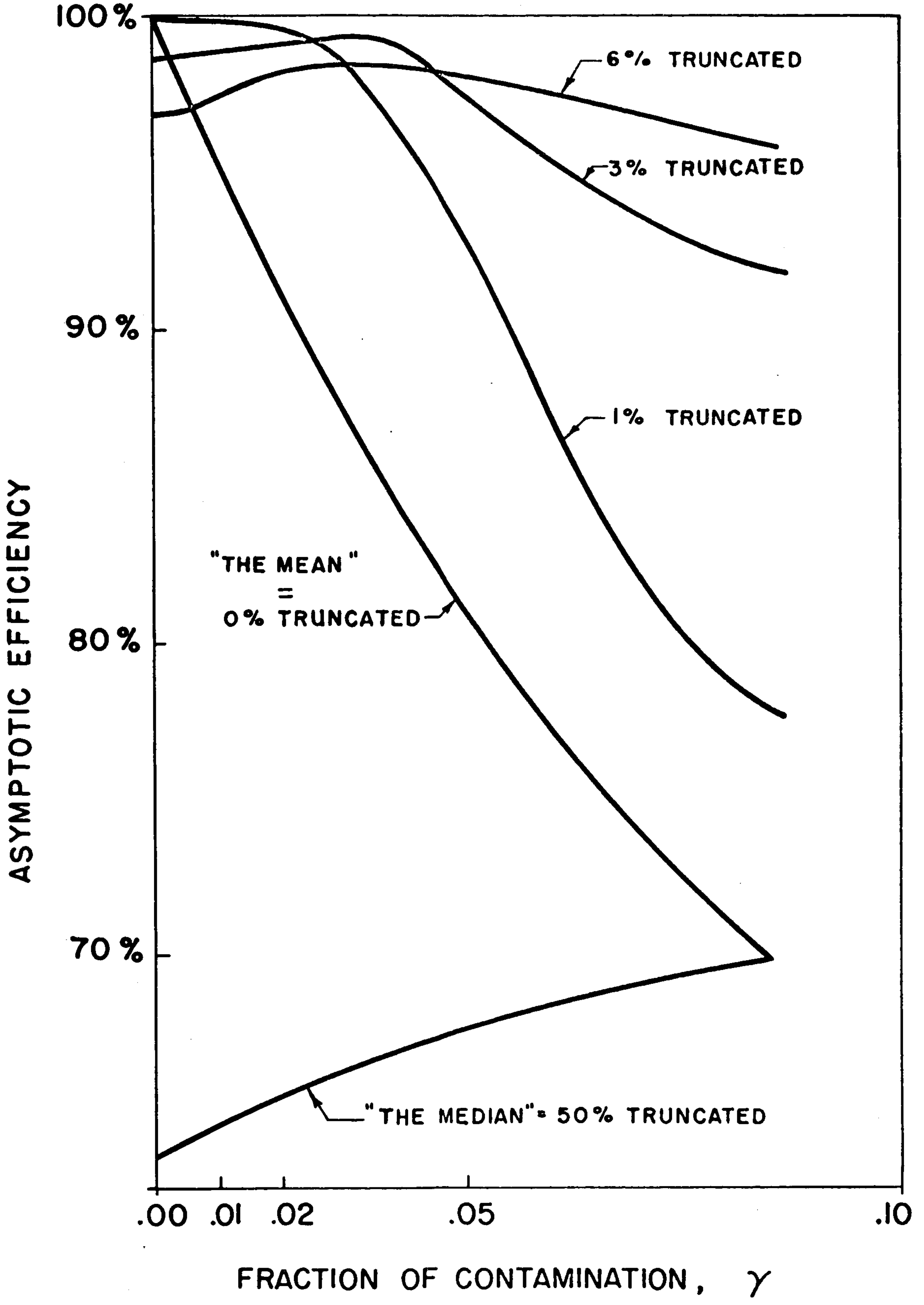


FIGURE 11 Asymptotic efficiency, for location, of truncated means.

REFERENCES

1. Electro-Mechanical Research, Inc., Report, "Statistical Analysis of Simulated USB Static Doppler Tracking Data", September, 1966, Contract NAS 5-3743.

2. Blom, Gunnar, Statistical Estimates and Transformed Beta-Variables, John Wiley and Sons, Inc., New York, 1958.

3. Tukey, John W., "A Survey of Sampling from Contaminated Distribution", Paper 39 (pp. 448-485) in Contributions to Probability and Statistics (I. Olkin et al, eds.), Stanford University Press, 1960.

4. Bickel, P. J., "On Some Robust Estimates of Location", Ann. of Math. Stat. 36, 1965, pp. 847-858.

5. Huber, Peter J., "Robust Estimation of A Location Parameter", Ann. of Math. Stat. 35, 1964, pp. 73-101.

6. Tukey, John W., "The Future of Data Analysis", Ann. of Math. Stat. 33, 1962, pp. 1-67.

7. Willke, T. A., "A Note on Contaminated Samples of Size Three", Journal of Research of National Bureau of Standards 70 B, 1966, pp. 149-151.

8. Sarhan, Ahmed F. and Greenberg, Bernard G., (eds.), Contributions to Order Statistics, John Wiley and Sons, Inc., New York, 1962.